LEVEL II

Form Approved
Budget Bureau No. 22-R0293
March 1976

Final Report
Covering the Period 15 October 1974 through 1 December 1975
Stanford Research Institute Project 3843

ASSESSMENT OF ACCEPTABILITY OF DIGITAL SPEECH COMMUNICATION SYSTEMS

by

Richard W. Becker
Project Leader
(415) 326-6200, Ext. 4325

Contract DAHC04-75-C-0008
ARPA Order Number 2905
Program Code Number 62706E

Sponsored by

Defense Advanced Research Projects Agency
Arlington, Virginia 22209

Effective Date of Contract: 15 October 1974
Contract Expiration Date: 1 December 1975
Amount of Contract: $84,876.00

ABSTRACT

Intelligibility tests were conducted for speech presented over eight bandwidth compression systems. At a SNR of +26 dBA there is little difference in intelligibility scores for the LPC and Channel Vocoder, both systems scoring higher than the (CVSD) system but lower than the Analog Reference system. The best performance among the digitized compression systems is achieved by the Culler-Harrison and Lincoln Lab LPC systems operating at 3500 bits/second, but the Channel Vocoder operating at 2400 bits/second was less adversely affected by noise at the speech input.

Except for the CVSD systems, adding more noise significantly degrades performance. The LPC systems are particularly vulnerable, suggesting that these systems have been designed to maximize operation in a noise-free environment. Further research should be oriented toward improving performance in typical noise environments.

CONTENTS

ILLUSTRATIONS

TABLES

I INTRODUCTION

This report describes the technical progress accomplished on Research Contract DAHC04-75-C-0008. These activities include conducting intelligibility tests on eight bandwidth compression systems, constructing a demonstration tape with samples of the various systems, processing a Quality Study Tape by each of the systems, recording new intelligibility test tapes under several different conditions, conducting speaker recognition tests on several of the systems, developing an automatic tape matrixing program on a PDP-11/40 computer, developing a program to introduce delays into LPC-Vocoded speech, developing two consoles that can be used to test the effect of various time-domain communication distortions upon task performance, and cooperating in designing a quality testing program for the ARPA ALOHA project.

II INTELLIGIBILITY TESTS ON NINE SYSTEMS AT THREE SPEECH-TO-NOISE RATIOS

Intelligibility testing was done using a version of the Modified Rhyme Test (MRT) constructed at SRI. This test consists of six lists of 50 monosyllabic words spoken by one male who was monitored during recording so as to assure relative constant vocal effort throughout the 300 items. Each test item is embedded in the carrier phrase "You will mark the...please."

A. Procedure

Noise was mixed with the speech to yield speech-to-noise ratios (SNR) of +3 dBA, +8 dBA, and +26 dBA on dubs. The noise that was added to the speech was created by passing white noise through a third-octave shaping network with the characteristics shown in Table 1. The purpose of this network was to create noise with a spectrum representative of the long-term adult male speech spectrum. This shape is also typical of general office and other environmental noises and for this reason serves as a fairly realistic source-receiver masking noise. Through the cooperation of ARPA contractors (Lincoln Labs and Culler-Harrison) and the Defense Communications Agency, the MRT tests at the three SNRs were processed by eight bandwidth compression systems and by a high-quality analog reference system.

The systems were evaluated in two testing programs. Each program used eight young adult males with normal hearing as listeners. The processed speech was played to the listeners at a comfortable level (approximately 65 dBA) using an Ampex 440 tape recorder and TDH-39 headphones with NAF-48490-1 muffs. Each program consisted of several half-day sessions with frequent rest breaks in each session. The order of

Table 1

RELATIVE ONE-THIRD OCTAVE WEIGHTS FOR NOISE SHAPING

Center Frequency (KHz)	Weight (dB)
0.025	-12
0.0315	-11
0.040	-10
0.050	- 9
0.063	- 8
0.080	- 7
0.100	- 6
0.125	- 5
0.160	- 4
0.200	- 3
0.250	- 2
0.315	- 1
0.400	0
0.500	- 1
0.630	- 3
0.800	- 5
1.000	- 6
1.250	- 8
1.600	- 9
2.000	-10
2.500	-12
3.150	-14
4.000	-17
5.000	-19
6.300	-23
8.000	-27
10.000	-30
12.500	-33
16.000	-35
20.000	-35

presentation was counterbalanced across systems and test lists. For details of recording and presentation see Becker and Kryter [1975].

B. Results

Table 2 shows the mean intelligibility scores, the standard deviation of the scores among listeners, and the standard error of the mean for the nine systems under the three different SNR levels. Figure 1 portrays the mean intelligibility scores plotted against bit rate for the different systems and Figure 2 shows percent MRT intelligibility of the systems as a function of SNR as measured at the speech input to the system.

C. Discussion

Figure 1 shows that, at a SNR of +26 dBA, there is little difference in the intelligibility scores for the LPC and Channel Vocoder and that these systems score higher than the CVSD system but lower than the Analog Reference system. Note that while Figure 1 shows results for LPC and Channel Vocoder systems at two different bit rates each, it is probably not reasonable to interpolate or extrapolate from these data, especially for the Channel Vocoder. Unlike systems such as the CVSD system in which two different bit rates generally reflect only quantitative differences in representation of a signal--for example, by means of greater precision of quantization--the LPC and especially the Channel Vocoder systems can be modified in fundamental ways to provide optimum operation at a specific bit rate.

Note also in Figure 1 that the best performance among the digitized compression systems is generally achieved by the Culler-Harrison and Lincoln Lab LPC systems operating at 3,500 bits per second. However, the Channel Vocoder when operated at 2,400 bits per second (the bit rate for which it was designed) is less adversely affected by noise at the speech input than are the two LPC systems tested.

Table 2

INTELLIGIBILITY OF NINE SYSTEMS UNDER THREE SPEECH-TO-NOISE CONDITIONS

Lab	Type	Bit Rate	System ID		SNR = +26	SNR = +8	SNR = +3
SRI	Analog Reference		1	MEAN	96.7	84.4	76.3
				SIGMA	2.5	5.2	7.5
				S. ERROR	0.36	0.76	1.09
DCA	CVSD	16.0	9		82.3	74.0	70.1
					6.7	8.2	7.2
					0.98	1.20	1.15
DCA	CVSD	9.6	8		72.2	72.6	68.6
					9.1	7.9	6.1
					1.32	1.15	0.89
Culler-Harrison	LPC	3.5	6		87.4	62.9	56.2
					4.8	9.9	10.5
					0.71	1.46	1.54
Lincoln Labs	LPC	3.5	5		85.9	62.9	53.7
					5.3	8.7	9.4
					0.79	1.27	1.37
Lincoln Labs	LPC	2.4	4		83.7	54.7	41.1
					4.8	8.0	15.3
					0.71	1.17	2.24
Lincoln Labs	Chan. Voc. Exp	4.8	3		82.7	65.5	51.6
					5.2	6.2	8.3
					0.76	0.90	1.21
Lincoln Labs	Chan. Voc. Exp	2.4	2		85.7	64.7	50.8
					5.7	6.5	11.1
					0.83	0.94	1.62
DCA	Chan. Voc. HY?	2.4	7		81.2	57.4	47.1
					6.0	12.4	10.3
					0.88	1.81	1.5

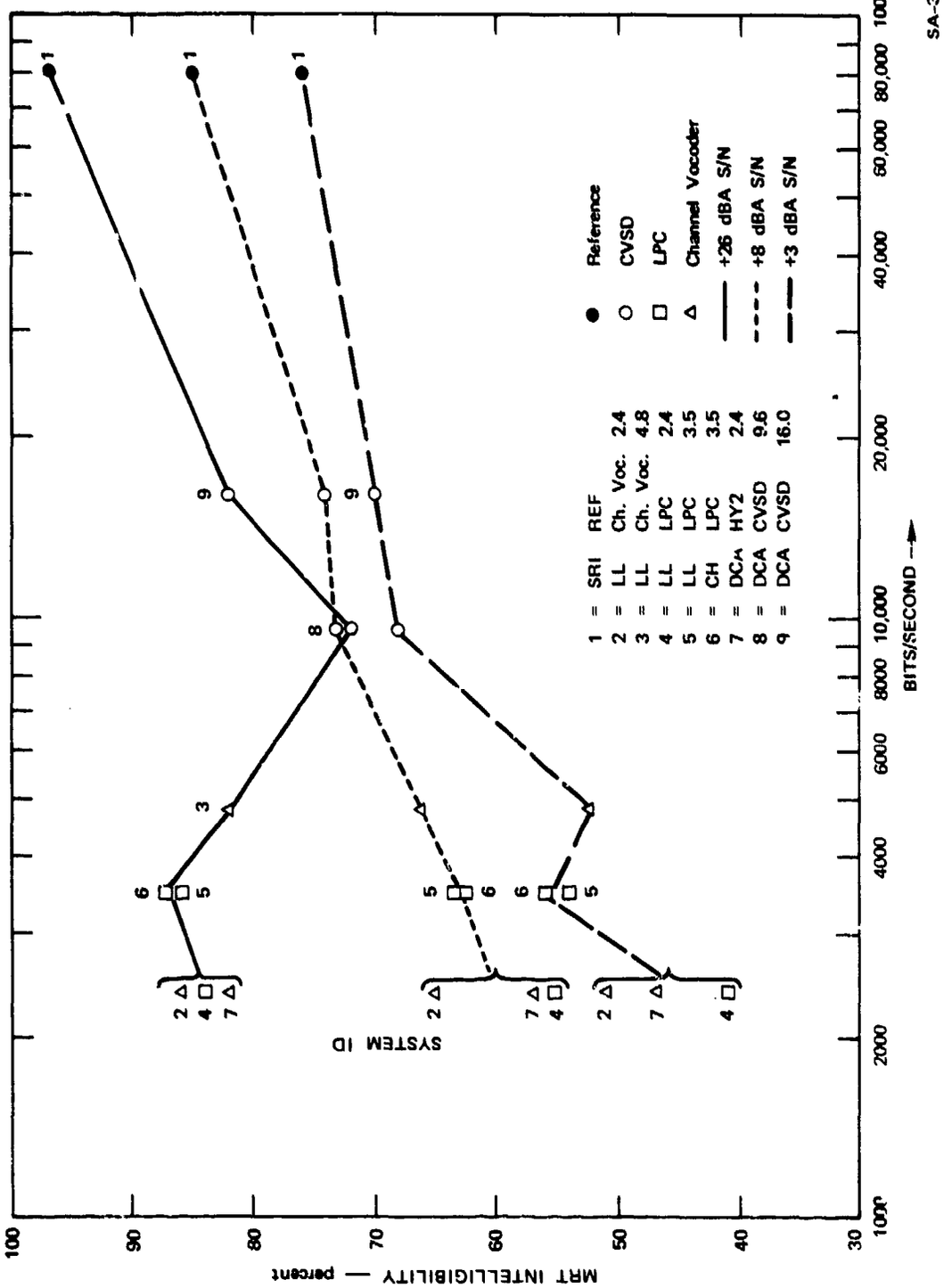

FIGURE 1 INTELLIGIBILITY AS A FUNCTION OF BIT RATE

The effects of the noise on the speech intelligibility scores are better shown in Figure 2. Examination of Figure 2 reveals that, except for the CVSD systems, adding more noise significantly degrades system performance and that the LPC systems are particularly vulnerable. While +20 dBA SNR is virtually equivalent to quiet in the case of the Analog Reference system, the informal listening tests reported in Section III suggest that in the case of the bandwidth compression systems even this small amount of added noise may result in degradation of intelligibility. It is possible that reduction of noise to a +40 dBA SNR could improve the intelligibility scores for systems such as the LPC systems.

The resistance of the CVSD systems to added noise is probably not surprising in that the degradation of intelligibility is probably caused by quantization noise approximately equivalent to +8 dBA SNR under the best of conditions. Thus, the addition of noise at the input simply results in degradation consistent with expectations of adding a small amount of noise to an already significant amount of noise. On the other hand, the LPC and Channel Vocoder systems show a degradation which suggest that they have been maximized for operation in noise-free environments, rather than in even moderate noise environments. Depending upon the proposed use of such systems within the Department of Defense, this might suggest that further research should be oriented toward improving performance in typical noise environments rather than in improving quiet performance.

The susceptibility of systems such as the LPC systems to noise raises questions concerning how the noise interacts with various parts of the bandwidth compression system. It has been standard procedure in intelligibility testing of communication systems to record the intelligibility test in an acoustically isolated room with a high quality microphone. This procedure results in master recordings that fully encompass

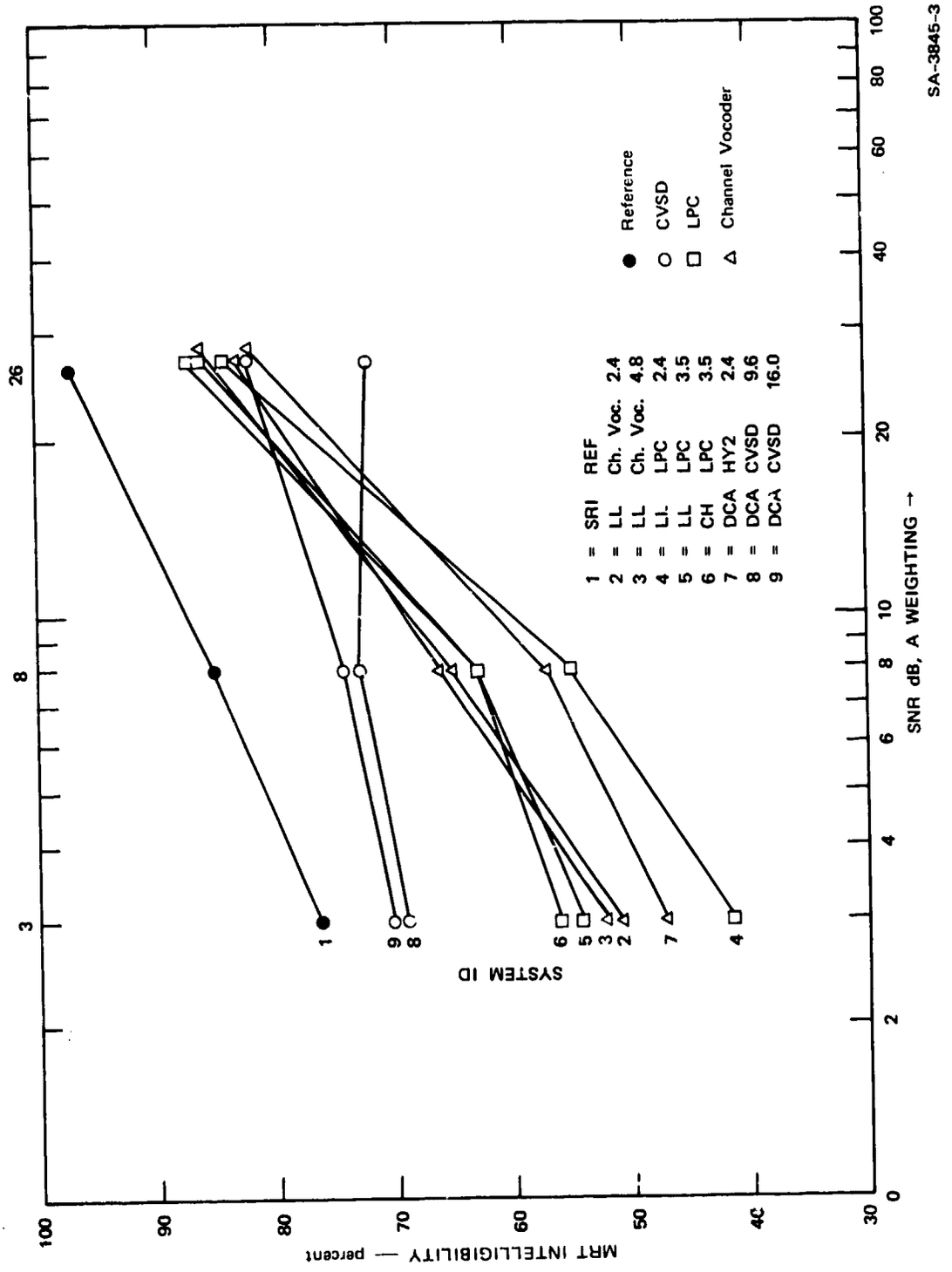

FIGURE 2 INTELLIGIBILITY AS A FUNCTION OF SNR

the 35 dB range of speech and the 10 kHz bandwidth of speech typically resulting in a peak speech-to-noise ratio in excess of +45 dB.

To test the performance of a system in noise, these recordings are typically electronically mixed with noise of specified loudness, bandwidth, and spectral shape and then processed by the system by putting the resultant recording into the electrical input of the system. Such a procedure ignores any effects that might be due to distortions introduced by the actual microphonic input in real system use. Most bandwidth compression systems that would be widely used would in fact rely upon the standard Western Electric 500 Telephone handset as the speech input device.

To accommodate this fact, we have recorded the Modified Rhyme Test of 300 words spoken under 3 different conditions. An experienced adult male speaker with a General American Dialect was used as the talker. In the first condition the 300 words (embedded in a carrier phrase) were spoken in "quiet" in a double-walled Industrial Acoustic Company sound isolation room. The sentences were recorded simultaneously with a B&K 1-inch condenser microphone on an Ampex 351 tape recorder and by a standard Western Electric handset on an Ampex 440 tape recorder. The telephone recording was also transmitted over a local exchange telephone line before recording. (Frequency characteristics of this link are available.)

The next condition was the same as the previous condition in terms of recording method, however, in this case the talker spoke the words while speech-shaped noise was played into the room at a level of 85 dBA. The resulting speech-to-noise ratio was approximately +8 dBA.

In the next condition, the noise was played into the room and recorded by both condenser mike and telephone handset without the talker speaking. This noise was then electronically mixed with the quiet recording of the words, again yielding a +8 dBA speech-to-noise ratio.

This set of recordings can be used to test the intelligibility performance of bandwidth compression systems that use common handsets for speech input. The recordings allow the separate investigation of effects due to the introduction of noise at the input, the effect of the processing of the noise by telephone before being processed, and the effects of increased vocal effort used by a talker in the presence of noise.

The administration of intelligibility tests in the evaluation of a large number of communication systems or communication conditions has traditionally been a very expensive, time-consuming task. The best known and most sensitive single-word intelligibility test is the Harvard Phonetically Balanced Word Lists test. However, accurate use of these lists requires extensive training of each crew of listeners so that they become very familiar with the entire 1,000-word vocabulary. Such training typically requires at least 12 hours per listener. After this training, the crew can typically be used for no more than 120 hours of listening before they have begun to learn the association of groups of words within each of the 20 lists and thus no longer give reliable results.

In an attempt to reduce the training time, a second set of tests have been developed that use a small, known, closed set for response alternatives. The best known of this kind of test is the Modified Rhyme Test (MRT). While this form of testing eliminates most of the training requirement in that overlearning of the vocabulary is no longer necessary, the length of time that a crew can be used is still limited because they eventually learn the word orders and associations within each of the lists.

Because of the considerable interlistener variance in intelligibility testing, it is desirable when comparing a set of communication

systems or conditions to use the same listeners for all conditions. The most effective way of accomplishigg this, while combatting the learning effects due to repeated listening to the same lists, is to constantly randomize the order of presentation of the items in each list. Unfortunately, making up new orders using the traditional reel-to-reel tape recorder dubbing method is time consuming and usually impractical in a program requiring hundreds of presentations of lists.

To solve this problem, we have designed and partially implemented an automatic matrixing system on the PDP-11/40 system at SRI. This system allows the storage of 50 utterances in digital format and the playing of the utterances for recording in any specified order. We expect that this system will greatly increase the efficiency of administering large-scale intelligibility testing programs.

III OTHER QUALITY MEASURES

A. Background

The total merit of a communication system must ultimately be measured by the degree to which the system transmits the same kinds and quantities of information that are communicated in an ideal communication system, such as face-to-face conversation in a quiet environment.

Some research has been conducted with the objective of measuring the quality of a communication system by using a single overall index of merit. Preference scaling is the major example of this approach [Munson and Karlin, 1962; Hecker and Williams, 1966]. Presumably, such unidimensional scales have proved of value in the comparative evaluation of systems, but they provide little information about the specific properties and capabilities of the systems so evaluated.

An alternative approach is to consider separately the different kinds of information that can be transmitted by speech. It seems unquestionable that the most significant information that must be transmitted is the semantic content of a message. The communication of this information has usually been tested by the use of single-word intelligibility tests. Tests that have been developed include Phonetically Balanced Word Lists (PBs) [Egan, 1948]; the Modified Rhyme Test (MRT) [House et al., 1965]; the Phonetically Balanced Rhyme Test (PBRT) [Clarke and Becker, 1969]; and the Diagnostic Rhyme Test (DRT) [Voiers, Sharpley, and Hehmsoth, 1973]. Each of these tests has advantages and disadvantages. For reasons of ease of administration, standardization, and widespread usage, the MRT has been used in our current research.

However, in addition to intelligibility, investigators have examined many other kinds of information that can be transmitted by a person's speech. A few of the attributes that have been investigated

include: the identity of the talker [Becker and Clarke, 1965; Pollack, Pickett, and Sumby, 1954]; the age of the talker [Ptacek and Sander, 1966]; the sex of the talker [Ingemann, 1968]; the personality of the talker [Markel, Eisler, and Reese, 1967]; and the emotional state of the talker [Fairbanks and Hoaglin, 1941]. Although listeners can do well in correctly identifying some of these traits from spoken utterances alone, performance tends to be rather poor on others. Nevertheless, some attributes have been identified with sufficient accuracy to warrant development of tests to measure how well a particular communication system transmits the properties of speech necessary to identify those attributes. Examples of attributes that have been incorporated in tests include: talker recognition and discrimination [Becker and Clarke, 1967]; semantic differential ratings of voices [Becker and Clarke, 1967; Voiers, 1964]; psychoacoustic ratings of voices [Becker and Clarke, 1967; Holmgren, 1963]; and emotional state as portrayed by actors [Becker and Clarke, 1967]. It has been found that, for analog transmission systems, the ability to transmit one kind of information has been highly correlated with the ability to transmit the other kinds of information. That is, if System A had a better ability to transmit single words intelligibly, it also had a greater ability to transmit the talker's identity, and so on. The advent of digital communication has changed this situation somewhat. Becker and Clarke [1967] reported a case in which a low-bandwidth system scores only 2 percent better on intelligibility than a Vocoder system while scoring 35 percent better on talker recognition. In our current program of research, similar results have been obtained.

B. Talker Recognition

Table 3 shows single word intelligibility scores and talker recognition scores for five digital bandwidth compression systems and an analog reference system. The intelligibility scores were obtained at SNRs of +26, +8, and +3 dBA. The talker recognition scores were obtained at +26 dBA and employed a name recognition test in which the voices of 16 talkers were processed by the systems and played to listeners who knew the talkers to determine if the listeners could identify the talkers. It can be seen that, while the CVSD 9.6 kbit system is inferior in intelligibility to the Channel Vocoders at +26 dBA SNR, it is definitely superior in conveying talker identification information. On the other hand, it can be seen that, while the CVSD system is inferior in talker recognition to the LPC systems at +26 dBA SNR, it is superior in intelligibility at the +8 and +3 dBA SNR conditions. These results suggest that, in evaluating the quality of speech transmitted over digital communication systems, one should measure more attributes of the speech than single-word intelligibility, and that these measures may vary independently with the speech-to-noise ratio.

While this kind of direct talker-recognition test has great validity, it suffers administratively in that it is often difficult to find a suitable group--of talkers and of listeners who know them--who are willing to participate in such a test. An alternative to this method is a talker discrimination test. One such test is a same-different test in which the voices of talkers are processed by a system, and listeners are then presented with pairs of voices that are either from the same talker or different talkers. The task of the listeners is to state whether the voices represent the same talker or different talkers. Because such a test can be administered to listeners who are not familiar with the talkers, it has much to recommend it. The test can

Table 3

PERCENTAGE INTELLIGIBILITY AND TALKER RECOGNITION
OF BANDWIDTH REDUCTION SYSTEMS

System	Percent Intelligibility SNR=26	SNR=8	SNR=3	Talker Recognition with SNR=26
Analog Reference	96%	84%	76%	89%
CVSD 9.6 kbit	72	72	68	72
LPC 3.5 kbit (CH)	87	63	56	83
LPC 3.5 kbit (LL)	86	63	54	78
EXP-Ch.Voc. 4.8 kbit	86	65	51	60
HY2-Ch.Voc. 2.4 kbit	81	57	47	58

be used to determine whether a given system transmits information that can be used to discriminate among talkers, but it has some disadvantages. Possibly, a communication system can transmit the information necessary for discrimination among talkers but also systematically change the attributes of talkers. That is, even though the talkers' voices may be so changed that people familiar with their natural voices would not recognize them, the systematic transformation of the voices may result in high discrimination scores on this same-different test.

For this reason, another kind of test has been constructed. The ABX test presents listeners with three voices: the voice of Talker A, the voice of Talker B, and the voice of Talker X, who is either A or B. The task of the listener is to state whether A or B is talking. If this test is administered by processing all three voices through a system to be tested, it suffers the same disadvantage as the same-different test. However, the test can also be administered in the form where the voices of Talker A and Talker B are not processed but the voice of X is. That is, the listener hears the natural voices of A and B, and then hears the voice of one of them after processing by the system being tested. The scores in such a test are much more indicative of the degree to which users of the system will be able to recognize talkers known to them, and also of the degree to which the system can be expected to sound natural to the users.

To develop such a test, we have recorded 160 adult males saying three sentences. The sentences are: "Pete Cooper's dog toyed with Dick Todd's cat," "But you and I should test fate," and "Shout if you sight the bird." All recordings were made in SRI offices with closed doors on a UHER 400 tape recorder using an Altec 633a dynamic microphone. One hundred twenty-nine of the same talker were recorded

saying the same three sentences again from 30 to 60 days later. After discarding some recordings because of the talker's speech pathology, we have 100 talkers saying each of three sentences at two different times separated by at least one month.

The automatic matrixing system can be used with these recordings to construct ABX tests including 300 items in which the same sentence is used for A, B, and X, and 900 items in which different sentences are used for A, B, and X.

C. Quality Study Tape

In the past, our experience in evaluating communication systems has suggested the desirability of subjecting new communication systems to a wide variety of audio environments and then using critical listening by experienced listeners to uncover potential problems with a system.

For these reasons, we have proceeded to develop a set of test materials that we have called a Quality Study Tape. This tape is intended for informal use to discover properties of systems that should be tested in a formal testing program. We have processed this tape over the systems described in Table 2, and our listening to date suggests several conclusions. First, when processing single words under excellent SNRs, most systems seem quite intelligible. Second, the systems seem to differ in their ability to transmit talker-identification information. Third, the presence of nonstationary background noise (in this case represented by a "typical" noisy office) had very different effects. The Channel Vocoders did not do well in transmitting the speaker's voice in this condition. The CVSD systems not only transmitted the speaker's voice well, but also accurately transmitted the character of the background noise (other people talking, a typewriter, and the like). The LPC systems did relatively well at transmitting the

main speaker's voice, although they occasionally lost it, and had an essentially suppressive effect upon the background noise. Fourth, each LPC system introduced very different characteristic background noise of its own. It is not clear which noise would be generally found more disturbing. Fifth, while all systems were highly intelligible when performing at high SNRs, there was a definite tendency of systems to systematically distort some phonemes, particularly in word final or sentence final position.

D. Demonstration Tape

As part of our informal evaluation of the quality of the bandwidth compression systems, we have constructed a demonstration tape. This tape, described in detail in Table 4, consists of selections of speech processed under various conditions from each of several representative bandwidth-compression systems. Copies of this tape have been made available to ARPA and authorized ARPA contractors.

Table 4

SYSTEM EVALUATION DEMONSTRATION TAPE I-L
(Full Track; 7-1/2 Inches/Sec.)

Order of Materials

1. P.J. saying "He ran five miles" with 2 different inflections in quiet.
2. R.M. saying "The purpose of this test is the investigation of interaction between talkers and communication systems" in quiet.
3. P.J. reading long passage in background of other conversations and office machine noise.
4. J.K. reading 5 MRT sentences. Speech-shaped noise added prior to processing to give +26 dBA SNR. Test words are: sing, book, nest, kith, and pun.
5. Same as 4 but SNR is +8 dBA.
6. Same as 4 but SNR is +3 dBA.

Order of Systems

1. High-quality Analog Reference system
2. Lincoln Labs Experimental Channel Vocoder--2.4 kbit
3. Lincoln Labs LPC system--3.5 kbit
4. Culler-Harrison LPC systems--3.5 kbit
5. CVSD System Recorded at Defense Communications Agency--9.6 kbit

IV EFFECTS OF TEMPORAL DOMAIN DISTORTIONS ON SPEECH QUALITY

A. Background

A packet communication network breaks up a message into small packets (typically on the order of 1,000 bits) and sends the packets from a source to a destination, typically through a series of repeater stations, to be reassembled into the message by the destination station. Error-free transmission can be accomplished through error-checking coding and retransmission of packets that are found to have errors. Obviously, such a process may introduce delays into the communication process. These delays occur as the result of the initial collection of a packet before transmission; the time required to retransmit some packets; and, when the system is heavily loaded, the unavailability of routes. The effects of such delays on speech communication depend on the length and distribution of the delays, the reassembly strategies used at the destination station, and the kind of communication task.

In the evaluation of various transmission and reconstruction strategies, the major trade-off is between what we shall call communication lag and interruption rate. Communication lag is the length of time between when a message is originated by the speaker and when it is first heard by the listener. Two parameters are associated with this variable. One is the duration between the beginning of the spoken message and the beginning of the heard message; the other is the duration between the end of the spoken message and the end of the heard message. The values of these parameters will depend on the transmission and reconstruction strategies used and on the configuration of the network during transmission. In general, the lags will not be constant but rather will be described by a probability distribution function with an expected value, variance, skewness, and so forth. In one extreme example of transmission and reconstruction strategies, a packet of 1,000 bits would simply be sent

as soon as it had been collected at the transmission terminal and be put out to the listener as soon as it had been collected at the receiving terminal. In this case, the mean communication lag between the beginnings of the transmitted and the received messages would be the mean time of transmission through the network under its current configuration and load. Under these circumstances, the expected value of the lag between end of spoken message and end of heard message would be the same.

If, however, we modify this transmission strategy so that a packet is not transmitted until verification has been received that the previous packet has been successfully transmitted and played out to the listener, then the mean lag between beginning of spoken message and beginning of heard message remains the mean time for transmission through the network. However, the expected time between end of spoken message and end of heard message could be much longer, depending on the time required for the successful transmission of each packet and verification of the transmission.

At the other extreme of transmission-reconstruction strategies would be the case in which a message would be collected in its entirety at the receiving terminal (its end being indicated by an end-of-message packet) before being played to the listener. In this extreme case, both the beginning-of-message lag and the end-of-message lag would be equal to the duration of the message plus the mean duration for transmission of a packet through the given network.

B. Problems of Temporal Distortion

The effect of communication lag on communication performance (by which we mean the efficiency, accuracy, and ease of usage) by humans depends on the nature of the communication task. At one extreme, a strictly one-way communication, such as leaving a message for someone to listen to later, the effect would be essentially zero. At the other extreme--two-way interchange of information requiring much feedback and

verification in a noisy environment--the effect would be considerable. We would expect the time needed to achieve an accurate interchange of information to be much greater than that for a standard telephone conversation simply because of the requirement that a message not be started at the listener end until its completion at the speaker end. Furthermore, we would expect this requirement to impose a considerable burden on those talking, forcing them into an unnatural mode of communication, with resulting loss of efficiency.

A second variable that affects communication performance is interruption rate. This variable is described by the durations of uninterrupted speech, together with the durations of the intervals between such speech. If, for example, a transmission-reconstruction strategy were used in which a packet was not sent until transmission of the previous packet had been finished, the listener would hear: a packet of speech, silence for an interval while the verification was being transmitted and the next packet was being transmitted, the next packet of speech, an interval of silence, and so on. Depending on the relative durations of the uninterrupted speech and of the intervals of silence, communication performance might be affected very little or be totally destroyed.

Because temporal distortion of this kind has not been characteristic of analog transmission systems, or even of circuit-switched digital communication systems, virtually no research has been conducted on the effects of such distortion on the intelligibility and quality of speech communications. However, certain related research suggests that the effects of this distortion are potentially very large.

For example, Miller and Licklider [1950] conducted experiments in which segments of an utterance were removed and replaced by silent intervals. They found that, depending on the percentage of speech they discarded and the frequency with which this was done, single-word

intelligibility in high-quality speech was decreased as much as 50 percent. Fairbanks, Guttman, and Miron [1957] found that, if small portions of the utterance were discarded and the remaining parts were joined rather than being separated by gaps, thereby increasing the speed of presentation of the utterance, comprehension of spoken passages was decreased. Discarding 60 percent of the passage resulted in a 50 percent decrease in comprehension. Other investigators [Schubert and Parker, 1955; Huggins, 1972] have found that, if a spoken passage is alternated from ear to ear, some rates of alternation greatly reduce the intelligibility of the utterance.

All this research shares the characteristic that the speech to a single ear did not contain the entire utterance; that is, parts of the utterance were actually missing. In addition, the Fairbanks study not only discarded part of the speech but also presented it at a more rapid rate than that at which it had been uttered, which would be expected to reduce comprehension. These studies are not representative of the conditions that are expected to occur in packet network communication. The temporal distortion in packet networks will typically consist of presenting the entire utterance to one or both ears, while only introducing gaps in the message rather than actually discarding parts of it.

It is reasonable to expect that, if we actually discard parts of a spoken utterance, single-word intelligibility will be degraded, since the discarded parts will potentially include whole phonemes or parts of words. One might not expect this effect if the entire utterance were presented even though interrupted by periods of silence. The only published reference to such a study is by Huggins [1972], in which he reported that intelligibility of connected speech can be destroyed by introducing 200-msec gaps every 20 msec of speech or by introducing 120-msec gaps every 62 msec of speech. These conclusions were verified by tests at SRI under a current research project [Becker and Kryter, 1976].

It should also be noted that even these experiments, which did not discard segments of speech but rather introduced gaps in the speech, were all done with a fixed amount of interrupted speech, a fixed interruption, and a fixed frequency of interruption. Thus none of these experiments represented the situation in which the mean durations of speech, mean durations of gaps, and mean frequency of interruption are specified but the actual values are varied according to some probability distribution.

In summary, evidence suggests that interruptions of speech may have disastrous effects on the intelligibility and quality of speech communication. Unfortunately, the values of interruption rate and interruption amount that have been shown to have significant effects on speech communication lie well within the boundaries of values that might be expected in some packet-switched networks.

Therefore, it is of critical importance to formally determine exactly how intelligibility and quality vary as a function of these parameters in order to recommend network strategies that will minimize undesirable effects. It is equally important to determine how these temporal distortion effects affect speech that is already degraded by bandwidth reduction processes, bad speech-to-noise ratios, low bandwidth, and so forth. No research in this area has been conducted, and such research is very important for determining strategies in the construction of future packet communication networks.

Even if interruptions and delays do not affect the intelligibility or other microaspects of quality, it is possible that such delays may seriously degrade two-way communication performance. One method of assessing communication performance over a particular system is to determine the accuracy or efficiency with which some specified task requiring two-way communication can be accomplished. Several such tasks

have been proposed. For example, Richards and Swaffield [1959] compared the time required by a listener to identify a figure in their conversation test with the time required by the same listener to identify a similar figure over a high-quality system. They found that this ratio was highly correlated with the estimated effort required to understand sentences, which, in turn, was correlated with single-word intelligibility. However, the only kind of degradation that was investigated was the masking of speech by various degrees of white noise. It is reasonable to believe that the distortions introduced by digital communication systems, particularly the temporal distortions introduced by packet systems, may degrade communication performance in ways that are not correlated with single-word intelligibility.

C. Task for Measuring Communication Efficiency

Although several different tasks have been suggested for testing communication efficiency, no standardization has occurred in this area. In our current research program we have constructed two consoles that can be used for a variety of communication tasks. These consoles and tasks were adapted from the work of Cohen et al. [1973], who used the consoles to measure the effects of environmental noise on task performance. The consoles (represented schematically in Figure 3) consist of a vertical array of four lights and a horizontal array of four digits.

The displays on these consoles are controlled by a PDP-11/10 computer, which can also read the response of each subject and measure the time required to make the response. The tasks that can be performed on these consoles vary from quite simple to very complex. A typical experiment will require Subject 1 to inform Subject 2 which of his vertical lights is lit. Subject 2 will then tell Subject 1 what digit (1 through 4) corresponds to that position on his console. Subject 2

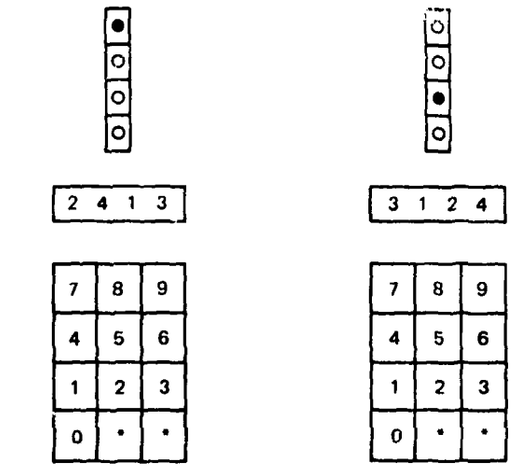

FIGURE 3 SCHEMATIC REPRESENTATION OF CONSOLES USED IN INFORMATION EXCHANGE TASKS

will then tell Subject 1 which light is lit on his vertical display. Subject 1 will then tell Subject 2 what digit is in that position on his digital display. At this point, both subjects key in the sum of the two digits. The displays are then changed: New vertical lights are lit, and new sets of horizontal digits are displayed.

The main virtue of these tasks is that the amount of information that must be exchanged to complete them successfully can be precisely controlled. It is thus possible to systematically investigate communication performance as a function of both channel degradation and complexity of task, as measured by the amount of information that must be exchanged.

The consoles will be used in future work to evaluate communication systems in which performance may be degraded by temporal domain distortions.

V ALOHA SPEECH EVALUATION

In cooperation with the ARPA ALOHA Satellite Program, we have recommended a series of quality tests for a combined speech and data transmission program. This program will employ iso-preference quality tests and Harvard sentence intelligibility tests for evaluation as recommended by this laboratory.

VI CONCLUSIONS

(1) Linear Predictive Coding (LPC) bandwidth compression systems operating at 3.5 kilobits/second provide significantly higher intelligibility than traditional 2.4 kilobit Channel Vocoders at both high and low speech-to-noise (SNR) ratios.

(2) LPC systems provide higher intelligibility than medium bandwidth systems such as Continuously Variable Slope Delta Modulation (CVSD) systems (9.6 kbit and 16.0 kbit) at high SNR ratios, but less intelligibility at low SNR ratios (+3 dBA to +8 dBA).

(3) Formal talker recognition testing and informal quality tests indicate that LPC systems convey significantly more parameters of talker recognition and naturalness than do either Channel Vocoders or CVSD systems.

(4) Temporal domain distortions of speech that might be encountered in packet-switched networks could significantly degrade the microaspects of communication such as intelligibility or talker discrimination, the croaspects of communication such as two-way communication efficiency, or both.

VII RECOMMENDATIONS

(1) Temporal domain distortions of speech will plague packet-switched networks until either very low bandwidth compression systems or very high bandwidth transmission networks are achieved. Such effects will only be exacerbated by attempts to add more people to the network or to add more people to a communications link as would be done, for example, in a conference situation. Little research has been done on the effects of time domain distortions upon intelligibility, naturalness, and overall communication performance. Therefore, we recommend that a comprehensive study be made of the effects of such temporal domain distortions upon the quality of speech communication.

(2) Temporal distortion may interact with other distortion products due to transducers, environmental noise, and so on. We recommend a comprehensive program to determine the effects of such interaction on the quality of human communication performance.

(3) LPC systems are very good at high SNRs. They are better both in intelligibility and naturalness than medium bandwidth systems, such as CVSD systems, even though the latter require three to four times as much bandwidth. This advantage is lost under conditions of low SNRs. We suggest that current LPC systems need to be developed further so as to achieve better performance in low SNR conditions. We suggest that research be conducted to determine:

(a) Can highly desirable intelligibility scores and naturalness be maintained under current bit rates even in high environmental noise by changing bits per parameter etc.

(b) If not, what increase in bit rate is needed to retain the good qualities of low bit rate LPC speech.

VIII PARTICIPATING SCIENTIFIC PERSONNEL

The following people participated in the research activities described in this report:

Richard W. Becker, Senior Research Psychologist

Karl D. Kryter, Staff Scientist

Earl J. Craighill, Research Engineer

James R. Young, Manager, Sensory Sciences Program

Fausto Poza, Senior Research Engineer

Donald W. Bell, Research Psychologist

James E. Davis, Engineering Associate

REFERENCES

R. Becker and F. Clarke, "Measurement of Speech Quality," Contract DAAB 03-67-C-0070 (NSA), Stanford Research Institute, Menlo Park, California (November 1967).

R. Becker and K. Kryter, "Assessment of the Acceptability of Digital Speech Communication Systems," Contract DAHC04-75-C-0008 (ARPA), Stanford Research Institute, Menlo Park, California (May 1975).

R. Becker and K. Kryter, "Quality of Digital Speech Communication Systems," Contract DAHC04-75-C-0008 (ARPA), Stanford Research Institute, Menlo Park, California (January 1976).

F. Clarke et al., "Technique for Evaluation of Speech Systems," Contract DA 28-043 AMC-00227(E) (US Army Electronics Laboratory, Stanford Research Institute, Menlo Park, California (April 1965).

F. Clarke and R. Becker, "Comparison of Techniques for Discriminating among Talkers," J. Speech and Hearing Res. (December 1969).

H. Cohen et al., "Noise Effects, Arousal, and Human Information Processing Task Difficulty and Performance," Contract NGL-34-002-055 (NASA), North Carolina State University (March 1973).

H. Dudley, "The Vocoder," Bell Lab. Record, Vol. 17, pp. 122-126 (1939).

J. Egan, "Articulation Testing Methods," Laryngoscope (1948).

G. Fairbanks, N. Guttman, and M. Miron, "Auditory Comprehension in Relation to Listening Rate and Selective Verbal Redundancy," J. Speech and Hearing Res. (1957).

G. Fairbanks and L. Hoaglin, "An Experimental Study of the Durational Characteristics of the Voice during the Expesssion of Emotion," Speech Monographs (1941).

J. Flanagan, Speech Analysis, Synthesis, and Perception (Springer Verlag; New York, 1972).

M. Hecker and C. Williams, "Choice of Reference Conditions for Speech Preference Tests," J. Acoust. Soc. Am. (1955).

M. Hecker and C. Williams, Final Report Yearby I, National Security Agency (1965).

L. Holmgren, "Speaker Recognition," Contract AFCRL-63-119 (May 1963).

A. S. House, et al., "Articulation Testing Methods: Consonantal Differentiation with a Closed Response Set," J. Acoust. Soc. Am. (1965).

A. House, C. Williams, M. Hecker, and K. Kryter, "Psychoacoustic Speech Tests: A Modified Rhyme Test," Tech. Doc. Rept. ESD-TDR-63-403, Stanford Research Institute, Menlo Park, California (June 1963).

A. Huggins, "Perception of Temporally Segmented Speech," Proc. Seventh Intl. Congress of Phonetic Sciences (1972).

F. Ingeman, "Identification of the Speaker's Sex from Voiceless Fricatives," J. Acoust. Soc. (1968).

R. Krauss and P. Bricker, "Effects of Transmission Delay and Access Delay on the Efficiency of Verbal Communication," J. Acoust. Soc. Am. (February 1967).

E. Kreul et al., "A Proposed Clinical Test of Speech Discrimination," J. Speech and Hearing Res., Vol. 11, pp. 536-548 (1968).

K. Kryter, K. Stevens, and M. Hecker, "An Evaluation of Speech Compression Systems," BBN report, Contract USAF 30(602)-2235 (RADC), (March 1962).

N. Markel, R. Eisler, and H. Reese, "Judging Personality from Dialect," J. Verbal Learning and Verbal Behavior (1967).

G. Miller and J. Licklider, "The Intelligibility of Interrupted Speech," J. Acoust. Soc. Am. (1950).

W. Munson and J. Karlin, "Isopreference Method for Evaluating Speech-Transmission Circuits," J. Acoust. Soc. Am., Vol. 34, pp. 762-774 (1962).

I. Pollack, J. Pickett, and W. Sumby, "On the Identification of Speakers by Voice," J. Acoust, Soc. Am., Vol. 26, p. 403 (1964).

R. Ptacek and E. Sander, "Age Recognition from Voice," J. Speech and Hearing Res., p. 273 (1966).

D. Richards and J. Swaffield, "Assessment of Speech Communication Links," Proc. Inst. Elec. Engrs., Vol. 106, Part B, No. 26, p. 77 (1959).

E. Schubert and C. Parker, "Addition to Cherry's Findings on Switching Speech Between the Two Ears," J. Acoust. Soc. Am. (1955).

W. Voiers, "Perceptual Bases of Speaker Identity," J. Acoust. Soc. Am., Vol. 36, p. 1065 (1964).

W. Voiers, A. Sharpley, and C. Hehmsoth, "Research on Diagnostic Evaluation of Speech Intelligibility," Tracor report, Contract F19628-70-C-0182 (January 1973).